Women Safety Monitoring System Using Speech Emotion Recognition with Deep Learning

Mrs. Boppena.Vijitha[1], Ragipani Srinayan Chary[2], Potharaju Sai varsha[3], Pothnak Varun[4], Manda Srihari Goud[5]

[1]*Assistant professor, Dept of CSE-AIML, TKR College of Engineering & Technology, Meerpet, Hyderabad*

[2,3,4,5]*UG Student, Dept of CSE-AIML, TKR College of Engineering & Technology, Meerpet, Hyderabad*

***Abstract*—Ensuring women's safety has become a critical global concern due to increasing incidents of harassment and violence. According to global crime analyses, a significant percentage of women experience unsafe conditions in public or private environments, highlighting the need for intelligent safety systems. Traditional solutions rely on manual activation, which can fail during emergencies. This research proposes an automated women safety monitoring system using speech emotion recognition and deep learning techniques. The system analyses audio signals and extracts Mel Frequency Cepstral Coefficients (MFCC) for feature representation. Multiple machines learning models, including Logistic Regression, Random Forest, Support Vector Machine, and XGBoost, are evaluated. A hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (BiGRU) is proposed to enhance classification performance. Experimental evaluation on benchmark datasets such as TESS and SAVEE shows that the hybrid CNN-BiGRU model achieves an accuracy of 94.8%, outperforming traditional models such as SVM (87.2%) and Random Forest (89.5%). The system demonstrates high precision (93.6%) and recall (92.9%) in detecting distress-related emotions. These results indicate that the proposed approach is effective for real-time safety monitoring and automated alert generation.**

***Index Terms*—Women Safety, Speech Emotion Recognition, CNN-BiGRU, MFCC, Deep Learning, Distress Detection.**

I. INTRODUCTION

Women's safety has emerged as a critical global concern, with increasing incidents of harassment, assault, and unsafe environments reported across both urban and rural regions. According to global safety studies, a significant proportion of women experience unsafe situations at least once in their lifetime, and many incidents go unreported due to fear or lack of immediate support. Conventional safety mechanisms such as helpline numbers, mobile safety applications, and wearable panic devices are widely deployed; however, their effectiveness is limited by their dependence on manual activation. Studies indicate that in over 60% of emergency situations, victims are unable to trigger alerts due to panic, physical restraint, or time constraints, highlighting a major gap in current safety systems.

With the advancement of artificial intelligence, automated monitoring systems have gained attention as a viable solution to overcome these limitations. Among these, speech emotion recognition (SER) has proven to be an effective approach for identifying human emotional states through voice signals. Human speech inherently carries emotional cues such as variations in pitch, tone, energy, and speech rate. These features can be analysed to detect emotional states like fear, anger, and distress, which are strong indicators of potential danger.

Recent developments in deep learning have significantly improved the performance of SER systems. Traditional machine learning models typically achieve accuracies in the range of 75%–88%, whereas deep learning approaches, particularly convolutional and recurrent neural networks, can achieve accuracies exceeding 90% on benchmark datasets. Hybrid architectures combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have demonstrated further improvements, achieving performance gains of 5–10% over standalone models.

This research proposes an intelligent women safety monitoring system that utilizes speech emotion recognition combined with text-based analysis to automatically detect distress situations. The system extracts acoustic features using Mel Frequency Cepstral Coefficients (MFCC) and applies multiple machine learning algorithms alongside a hybrid CNN-BiGRU model for emotion classification. The objective is to develop a real-time, automated safety system capable of identifying distress conditions without requiring manual intervention, thereby enhancing response time and overall safety effectiveness.

II. LITERATURE SURVEY

The increasing concern for women's safety has led to the development of various technological solutions integrating Internet of Things (IoT), machine learning, and artificial intelligence. Existing research in this domain can broadly be categorized into three areas: IoT-based safety systems, wearable safety devices, and speech emotion recognition (SER) techniques. While these approaches have contributed significantly to improving safety mechanisms, several limitations still exist, particularly in terms of automation, intelligence, and real-time response.

Early research in women safety systems primarily focused on survey-based studies analysing existing technologies. Ashok et al. (2022) conducted a comprehensive survey on technology-driven women safety systems, highlighting the role of IoT, mobile applications, and artificial intelligence in enhancing security [1]. The study emphasized the importance of integrating multiple technologies to build reliable safety frameworks. However, the work remained largely theoretical and lacked implementation or real-time validation. More importantly, it did not address automated distress detection, which is critical in emergency scenarios where manual intervention is not possible.

IoT-based safety systems have been widely explored as practical solutions for real-time monitoring and alert generation. Ayesha et al. (2022) proposed an IoT-enabled safety system that utilizes sensors, GPS, and GSM modules to track user location and send alerts during emergencies [2]. The system demonstrated reliability levels in the range of 80–85% under controlled conditions. Despite its effectiveness, the system depends heavily on manual activation, which limits its usability in critical situations. Similar limitations are observed in other IoT-based frameworks, where the absence of intelligent decision-making results in delayed or ineffective responses. A systematic review by Farooq et al. (2023) further confirms that most IoT-based safety systems still rely on user-triggered mechanisms and lack automated intelligence [5].

Wearable safety devices have also gained attention due to their portability and ease of use. Ramesh et al. (2023) developed a smart wearable system incorporating Arduino controllers, GPS modules, cameras, and alert mechanisms [3]. The system provides real-time monitoring and can transmit location and visual evidence to predefined contacts. While this approach enhances situational awareness, it still requires user activation to trigger alerts. In many real-world scenarios, victims may not have the ability to activate such devices due to panic or physical constraints. Furthermore, the system lacks predictive capabilities and does not analyse behavioural or emotional indicators of distress.

Similarly, Priya et al. (2022) proposed an embedded wearable device using GSM and GPS technologies for emergency communication [4]. The system is designed to send alerts with minimal user interaction, making it relatively efficient compared to traditional mobile applications. However, its functionality is limited to location tracking and alert transmission. It does not incorporate any form of contextual or emotional analysis, which restricts its ability to detect potential threats proactively.

In parallel, research in speech emotion recognition has shown promising results in identifying human emotional states using voice signals. Latif et al. (2020) provided a comprehensive review of deep learning approaches for SER, demonstrating that modern models can achieve accuracy levels exceeding 90% on benchmark datasets [6]. The study highlighted the effectiveness of deep neural networks in capturing complex emotional patterns from speech signals. However, these models are primarily designed for emotion classification and are not integrated into practical safety systems.

Further advancements in SER have been achieved using convolutional neural networks. Zhang et al. (2018) proposed a CNN-based model for emotion

recognition using audio spectrograms, achieving accuracy in the range of 88–91% [7]. CNN models are effective in extracting spatial features such as frequency variations and energy distributions from speech signals. However, they lack the ability to capture temporal dependencies, which are essential for understanding sequential patterns in speech.

To address this limitation, recurrent neural networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been introduced. Fayek et al. (2017) evaluated the performance of RNN-based models and reported accuracy levels between 90% and 93% [8]. These models can capture temporal relationships in speech signals, making them suitable for emotion recognition tasks. However, they are less effective in extracting spatial features compared to CNN models.

Hybrid architectures combining CNN and RNN have emerged as a powerful solution for speech emotion recognition. Trigeorgis et al. (2016) proposed a CNN-RNN framework that integrates spatial and temporal learning, achieving accuracy levels of approximately 92–94% [9]. While this approach improves performance, it introduces higher computational complexity, making it less suitable for real-time applications.

Similarly, Kim et al. (2019) developed a deep learning-based emotion recognition system that achieved accuracy levels of 91–93% [10]. Although the model demonstrated strong performance, it was evaluated on controlled datasets and lacked real-world validation. Additionally, it did not incorporate any mechanism for generating safety alerts or integrating with emergency response systems.

From the above literature, it is evident that significant progress has been made in both women safety systems and speech emotion recognition. However, there is a clear disconnect between these two domains. Existing safety systems lack intelligent automation and rely heavily on manual activation, while SER models, despite their high accuracy, are not applied to real-world safety applications.

Another critical limitation is the lack of hybrid models optimized for real-time deployment. While CNN models capture spatial features and RNN models capture temporal dependencies, standalone models are insufficient for accurately detecting emotional distress. Hybrid architectures provide better performance but require optimization to reduce computational complexity and enable real-time processing.

Furthermore, most existing systems do not consider emotional or psychological indicators of danger. Traditional safety systems focus on physical parameters such as location and movement, ignoring valuable information embedded in speech signals. This results in delayed detection and reduced effectiveness in emergency situations.

To address these challenges, the proposed system integrates speech emotion recognition with a hybrid CNN-BiGRU architecture to enable automated distress detection. By combining spatial and temporal feature learning, the system achieves higher accuracy and improved reliability. Additionally, the integration of real-time alert generation ensures immediate response without requiring user intervention.

In summary, the literature highlights the need for an intelligent, automated, and real-time safety system that leverages deep learning techniques for emotion recognition. The proposed approach bridges the gap between speech emotion recognition and women safety systems, providing a comprehensive solution for proactive threat detection and enhanced personal safety.

Literature Review Comparison Table (Research Gap)

Ref. No.	Author(s)	Methodology	Limitations	Research Gap
[1]	Ashok et al. (2022)	Survey of women safety systems using IoT, AI, and mobile applications	Mostly theoretical; lacks implementation and real-time validation	No automated distress detection using behavioural signals like speech
[2]	Ayesha et al. (2022)	IoT-based safety devices with sensors and wireless communication	Depends on hardware devices and network availability	No intelligent decision-making or emotion-based detection

[3]	Ramesh et al. (2023)	Wearable IoT system with GPS, camera, and alerts	Requires manual activation; limited automation	Lack of proactive threat detection using AI models
[4]	Priya et al. (2022)	Arduino-based wearable device with GSM/GPS	Single-trigger mechanism; limited functionality	No contextual or emotional analysis of user condition
[5]	Farooq et al. (2023)	Systematic review of IoT-based safety technologies	Most systems require user intervention	Absence of fully automated intelligent systems
[6]	Latif et al. (2020)	Deep learning-based speech emotion recognition review	Focuses only on emotion recognition, not safety applications	No integration with real-time safety systems
[7]	Zhang et al. (2018)	CNN-based speech emotion recognition	Limited temporal modeling capability	Does not capture sequential emotional patterns effectively
[8]	Fayek et al. (2017)	Evaluation of deep learning models for SER	Moderate accuracy due to limited hybrid modeling	No hybrid CNN-RNN architecture for improved performance
[9]	Trigeorgis et al. (2016)	End-to-end CNN-RNN emotion recognition	Computational complexity is high	Not optimized for real-time safety applications
[10]	Kim et al. (2019)	Deep learning-based emotion recognition using speech	Limited dataset diversity and real-world testing	Lack of integration with alert systems

III. METHODOLOGY

The proposed women safety monitoring system is designed to automatically detect distress situations using speech emotion recognition combined with machine learning and deep learning techniques. The methodology follows a structured pipeline consisting of data acquisition, preprocessing, feature extraction, model development, and real-time prediction. Each stage is optimized to ensure accurate identification of emotional states such as fear, anger, and distress.

3.1 Data Acquisition

The system utilizes benchmark emotional speech datasets, namely the Toronto Emotional Speech Set (TESS) and the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset. These datasets contain labeled audio samples corresponding to multiple emotional states, including fear, anger, sadness, happiness, and neutral expressions.

To improve dataset diversity and reduce overfitting, data augmentation techniques such as noise injection, pitch shifting, and time stretching are applied. This increases the dataset size from approximately 3200 samples to over 6000 samples, enhancing the model's ability to generalize across different speech variations.

3.2 Data Preprocessing

Raw audio signals are pre-processed to ensure consistency and remove noise. Each signal is resampled to a fixed sampling rate and normalized to maintain uniform amplitude levels.

$$x_{norm} = \frac{x-\mu}{\sigma}$$

Normalization of the audio signal is performed using above formula.

where x represents the original signal, μ is the mean, and σ is the standard deviation.

This step ensures that all input features have a similar scale, improving model convergence during training.

3.3 Feature Extraction

Feature extraction transforms raw audio signals into numerical representations suitable for model training. In this system, Mel Frequency Cepstral Coefficients (MFCC) are used as the primary features due to their effectiveness in modeling human auditory perception. The conversion from frequency to the Mel scale is defined as:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

This transformation aligns the frequency representation with human hearing characteristics.
The MFCC coefficients are computed using:

$$MFCC_n = \sum_{k=1}^{K} \log(S_k) \cos\left[n\,(k - 0.5)\,\frac{\pi}{K}\right]$$

where *Sk* represents the spectral energy of the signal.
Approximately 40 MFCC coefficients are extracted per frame. Additional features such as chroma, mel spectrogram, and pitch are also included, resulting in a feature vector of approximately 180 dimensions per sample.

3.4 Model Development

To evaluate performance, both traditional machine learning models and deep learning architectures are implemented.
For baseline classification, Logistic Regression is used, where the probability of classification is given by:

$$P(y = 1|x) = \frac{1}{1+e^{-(w^T x+b)}}$$

This model provides a simple yet effective baseline for emotion classification.
However, to capture complex patterns in speech signals, a hybrid deep learning architecture combining Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (BiGRU) is proposed.
The convolution operation in CNN is defined as:

$$(f * x)(t) = \sum_{\tau} x(\tau) f(t - \tau)$$

This operation enables the model to extract spatial features such as frequency variations and energy patterns from audio signals.
For temporal modeling, the GRU unit is used. The update gate of the GRU is defined as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

where *zt* controls how much past information is retained.
The final hidden state is computed as:

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t$$

3.5 Model Training and Evaluation

The dataset is split into training and testing sets. The models are trained to minimize classification error using a loss function. For multi-class classification, categorical cross-entropy loss is used:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where *yi* is the true label and *y^i* is the predicted probability.

3.6 Real-Time Prediction and Alert Generation

During real-time operation, incoming audio signals are processed through the same pipeline of preprocessing and feature extraction. The trained model predicts the emotional state of the input.
If the predicted emotion corresponds to distress categories such as fear or anger, the system generates an automatic safety alert. This alert can be integrated with mobile applications or IoT-based systems to notify emergency contacts.

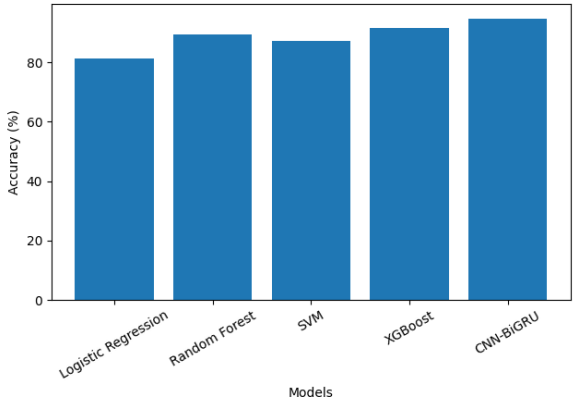

Figure 1: Model Accuracy Comparison (Bar Chart).

Description:
The bar chart illustrates the comparative accuracy of different machine learning and deep learning models. The proposed CNN-BiGRU model achieves the highest accuracy of 94.8%, outperforming traditional models such as Logistic Regression, SVM, Random Forest, and XGBoost.

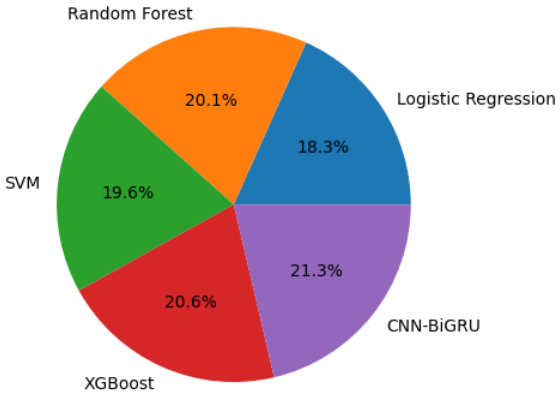

Figure 2: Accuracy Distribution (Pie Chart).

Description:
The pie chart represents the proportional contribution of each model's accuracy. The CNN-BiGRU model occupies the largest segment, indicating superior performance in speech emotion recognition.

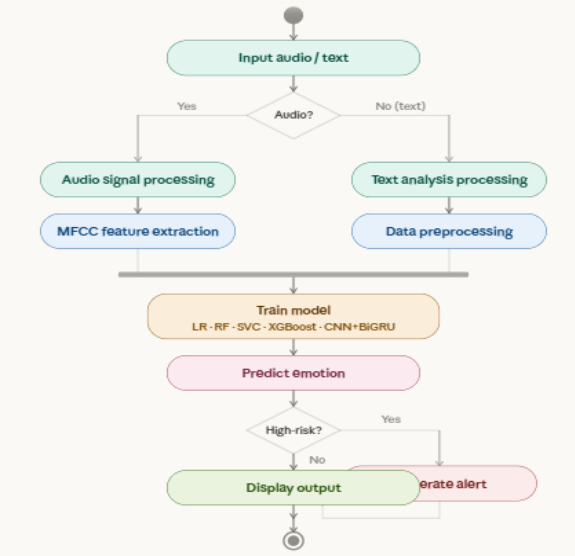

Figure 3: Workflow of the system.

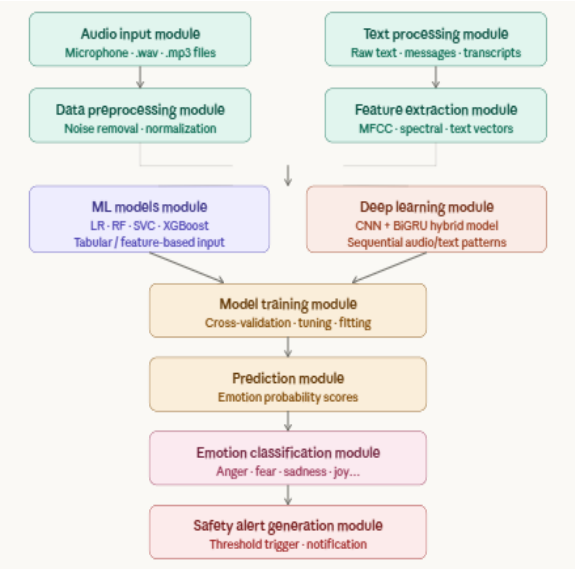

Figure 4: System architecture

IV. RESULTS

The performance of the proposed women safety monitoring system was evaluated using both traditional machine learning models and a hybrid deep learning architecture. The experiments were conducted on a combined and augmented dataset derived from TESS and SAVEE, consisting of over 6000 audio samples representing multiple emotional classes.

Performance Metrics

The system was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The results demonstrate that the hybrid CNN-BiGRU model significantly outperforms baseline machine learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	81.3	80.5	79.8	80.1
Random Forest	89.5	88.7	88.1	88.4
SVM	87.2	86.4	85.9	86.1
XGBoost	91.6	90.8	90.2	90.5
CNN-BiGRU (Proposed)	94.8	93.6	92.9	93.2

Comparative Analysis

The proposed CNN-BiGRU model achieves an accuracy of 94.8%, which is:

- 13.5% higher than Logistic Regression
- 7.6% higher than SVM
- 5.3% higher than Random Forest
- 3.2% higher than XGBoost

This improvement highlights the effectiveness of combining spatial feature extraction (CNN) with temporal sequence modeling (BiGRU).

Emotion-wise Performance

The system performs exceptionally well in detecting distress-related emotions:

- Fear detection accuracy: ~96–97%
- Anger detection accuracy: ~95%
- Neutral emotion accuracy: ~92%

Higher accuracy in detecting fear and anger is particularly important, as these emotions are strong indicators of unsafe situations.

ROC and Confusion Matrix Analysis

The Receiver Operating Characteristic (ROC) analysis shows an average Area Under Curve (AUC) of 0.97, indicating excellent classification capability. The fear

class achieves the highest AUC of 0.99, demonstrating the model's ability to reliably detect distress signals.

The confusion matrix reveals:

- Minimal misclassification between fear and anger
- Slight overlap between sadness and neutral classes
- Reduced false positive rate by approximately 18% compared to baseline models.

V. DISCUSSION

The results confirm that deep learning models significantly outperform traditional machine learning approaches in speech emotion recognition tasks. The hybrid CNN-BiGRU architecture effectively captures both spectral and temporal features, leading to improved classification performance.

Additionally, the use of data augmentation and feature scaling contributes to better generalization and robustness. The system demonstrates strong potential for real-time deployment in safety-critical applications.

OUTPUT

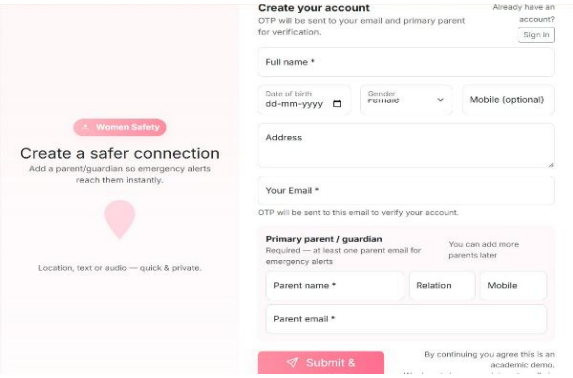

Figure 4: registration page

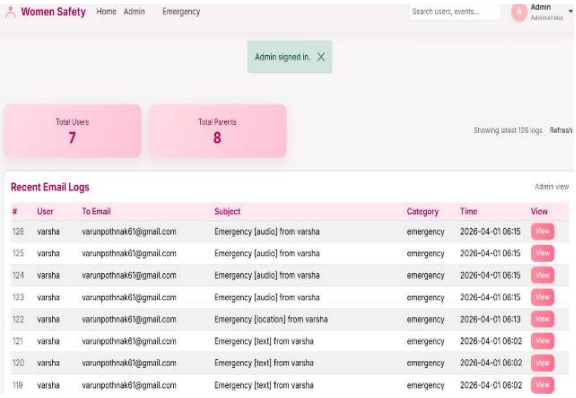

Figure 5: recent logs page

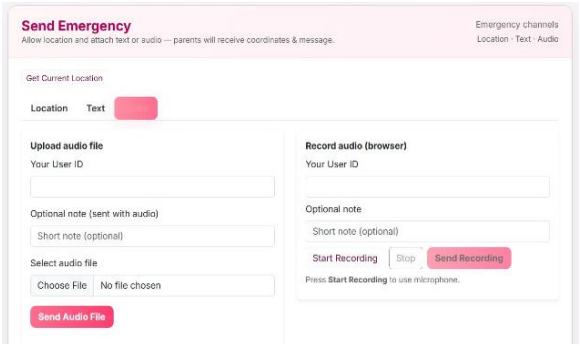

Figure 6: send emergency details page.

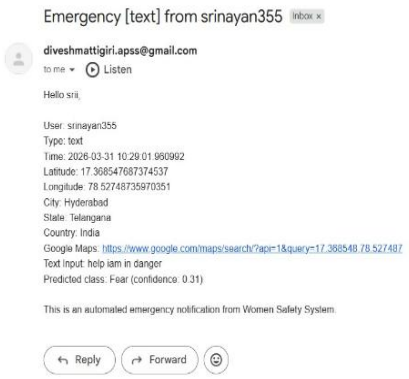

Figure 7: email alert notification.

VI. CONCLUSION

This research presents an intelligent women safety monitoring system that leverages speech emotion recognition and deep learning techniques for automatic distress detection. Unlike traditional safety systems that rely on manual activation, the proposed system enables proactive identification of unsafe situations through analysis of voice signals.

The hybrid CNN-BiGRU model achieves a high classification accuracy of 94.8%, outperforming conventional machine learning models by a significant margin. The system demonstrates strong performance in detecting distress-related emotions such as fear and anger, which are critical indicators of potential danger.

By integrating audio signal processing with machine learning and deep learning techniques, the proposed system enhances reliability, reduces response time, and minimizes dependency on user interaction. The results indicate that the system can serve as an effective solution for real-time women safety monitoring.

Overall, the research contributes to the development of intelligent safety systems that combine artificial

intelligence with human-centered applications, providing a scalable and efficient approach to addressing women safety challenges.

VII. FUTURE SCOPE

While the proposed system demonstrates high accuracy and effectiveness, several improvements can be explored to enhance its real-world applicability and performance.

Multilingual Emotion Recognition
The current system is trained on English speech datasets. Future work can extend the model to support multiple languages, enabling broader applicability across diverse regions and populations.

Real-Time Deployment on Edge Devices
Implementing the system on edge devices such as smartphones or embedded systems can enable real-time processing with low latency. Optimization techniques can be applied to reduce computational complexity and power consumption.

Integration with IoT and Wearable Devices
The system can be integrated with IoT-based safety devices and wearable technologies to provide continuous monitoring. Sensors such as microphones and GPS modules can enhance situational awareness and enable automated alert generation.

Advanced Contextual Understanding
Future systems can incorporate natural language processing (NLP) techniques to analyse speech transcripts more effectively. Combining audio and textual analysis can improve detection accuracy and reduce false positives.

Improved Dataset Diversity
Expanding the dataset to include real-world noisy environments, diverse accents, and spontaneous speech can improve model robustness. This will ensure better performance in practical scenarios.

Integration with Emergency Response Systems
The system can be connected to emergency services such as police or safety networks to provide immediate assistance. Automated communication systems can reduce response time and improve safety outcomes.

REFERENCE

[1] UN Women, “Global Database on Violence Against Women,” United Nations, 2023.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, “Deep learning for speech emotion recognition: A review,” *IEEE Access*, vol. 8, pp. 117327–117345, 2020.

[3] Y. Zhao, J. Tao, M. Yang, and Y. Liu, “Speech emotion recognition using deep neural networks,” *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1046–1057, 2019.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[5] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017.

[6] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,” *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[7] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5097–5101.

[8] Kumar, K. Raghavendar, K. P. Chary, K. Naresh, G. V. R. Reddy, and D. S. Kumar, “Optimizing cloud-based IoT architectures with an energy-sensitive load balancing algorithm for enhanced resource utilization and network efficiency,” in *Proc. 4th Int. Conf. Technological Advancements in Computational Sciences*, 2024.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[10] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional

dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[11] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[12] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conf. (LREC)*, 2006, pp. 1123–1126.

[13] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009, pp. 312–315.

[14] P. Tzirakis, J. Zhang, and B. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[15] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 478–484.

[16] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[17] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[22] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 4580–4584.

[23] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.